

Stopping rules for phase II studies

Nigel Stallard, John Whitehead, Susan Todd & Anne Whitehead

Medical and Pharmaceutical Statistics Research Unit, The University of Reading, PO Box 240, Earley Gate, Reading, Berkshire, RG6 6FN

This paper, the second in a series of three papers concerned with the statistical aspects of interim analyses in clinical trials, is concerned with stopping rules in phase II clinical trials. Phase II trials are generally small-scale studies, and may include one or more experimental treatments with or without a control. A common feature is that the results primarily determine the course of further clinical evaluation of a treatment rather than providing definitive evidence of treatment efficacy. This means that there is more flexibility available in the design and analysis of such studies than in phase III trials. This has led to a range of different approaches being taken to the statistical design of stopping rules for such trials. This paper briefly describes and compares the different approaches. In most cases the stopping rules can be described and implemented easily without knowledge of the detailed statistical and computational methods used to obtain the rules.

Keywords: clinical trial design, early stopping, group sequential designs

1. Background

This is the second in a series of three papers discussing the statistical aspects of interim analyses and early stopping in clinical trials. In the first paper, Todd *et al.* [1] discussed the use of interim analyses in phase III clinical trials. Working backwards through the process of clinical trials in drug development, this paper focuses on phase II trials. The third paper will consider phase I, dose-finding, studies.

The term *phase II clinical trial* is used to describe a wide variety of investigative studies. These range from very small single-arm studies to assess treatment efficacy and safety of an experimental treatment prior to phase III evaluation, to randomised studies comparing a control treatment with the experimental therapy at several doses. Common features of phase II trials as opposed to phase I studies are the use of a patient population rather than healthy volunteers and the assessment of treatment efficacy as well as safety. In contrast to phase III studies, phase II trials are not primarily designed to give definitive evidence of treatment efficacy, the outcome of the trial being further (phase III) testing rather than regulatory submission

or direct influence on clinical practice. Interpretation of phase II trials is therefore generally more informal than for phase III trials, and often leads only to internal decision making within a pharmaceutical company or medical research institute.

Because phase II trials are usually small and relatively short-term, attention is generally focused on rapidly observable responses, facilitating monitoring of the data as they accumulate. In serious diseases, the data will be monitored and the trial stopped either if there is evidence of a lack of safety or efficacy associated with the experimental therapy or if there is sufficient evidence of efficacy to warrant phase III testing. Whether or not it is specified in advance, the conduct of the phase II trial may thus be described by a *stopping rule* setting out the circumstances under which the trial will end and the action that will then be taken. An appropriate stopping rule can lead quickly to an indication of whether the treatment is efficacious, and a decision to start confirmatory phase III trials based on longer-term endpoints such as survival times. To date, most of the development of stopping rules for phase II trials has been motivated by trials in serious diseases, particularly cancer. With appropriate modifications, however, the resulting methods should also lead to efficiency advantages in less serious conditions.

This paper gives a brief description of some of the approaches that have been suggested for the choice of stopping rules for phase II studies. As phase II trials vary

Correspondence: Dr. N. Stallard, Medical and Pharmaceutical Statistics Research Unit, The University of Reading, PO Box 240, Earley Gate, Reading, Berkshire, RG6 6FN. Tel.: 0118 9316565; Fax: 0118 9753169; E-mail: n.stallard@reading.ac.uk

Received 18 April 2000, accepted 10 February 2001.

considerably in their size and objectives, these approaches are diverse. Some trials are conducted like small-scale phase III trials, in which case methods that parallel those in phase III are suitable: careful adjustment to the final analysis has to be made to allow for the monitoring of the trial. In many phase II trials, it is not appropriate to summarize results by a hypothesis test with specified error rates. It is more important that the trial can be shown to have satisfactory properties in terms of the probability that effective treatments proceed to phase III while ineffective treatments do not. If the hypothesis test formulation is abandoned, the associated adjustment of error rates to allow for interim analyses ceases to be relevant.

Phase II trials can be divided into three broad categories: *single-arm studies*, in which all patients receive the same experimental treatment at the same dose, *comparative studies* in which patients are randomised between a single experimental treatment and a control treatment, and *selection studies*, in which a number of different treatments or doses are compared, possibly without a control group. As the statistical aspects of these types of trials are different, they will be considered separately in sections 2–4 below.

Although phase II studies are a common part of the drug development process in almost all therapeutic areas, the formal development of stopping rules has been particularly prominent in oncology. For this reason, this paper, which in part is a review of existing work, will tend to focus on cancer studies. As explained below, these usually do not include a control treatment. The data collected in such trials are usually in the form of success/failure outcomes. Although relatively little methodology currently exists, the ideas used in the construction of stopping rules can be extended to continuous and other types of response and to trials with an active or placebo control treatment.

2 Single-arm studies

In oncology or other serious diseases, phase II trials often include patients who have previously received unsuccessful treatment with standard therapies, so that all patients in a trial receive the experimental treatment. Although single-arm phase II trials are uncommon outside of oncology, this setting will be considered for two reasons. First, much of the recent work on stopping rule construction has been developed for this type of study. Second, as these trials are the simplest of all phase II trials, concepts can be illustrated most easily in the single-arm case prior to generalization.

If a success/failure response is used, the data from a single-arm trial can be summarized very simply as the number of patients included and the proportion of these for whom the treatment was successful. Although all patients receive the same treatment, such trials are intrinsically comparative in nature. In order to decide

whether further (phase III) testing is appropriate, it is necessary to assess the new treatment in comparison with either the known properties of the current standard treatment or some required level of activity.

The different approaches to phase II trial design which are described in this section will be illustrated by the design of a single-arm phase II trial described by Thall & Simon [2]. The purpose of the trial was to assess treatment with fludarabine + ara-C + granulocyte colony stimulating factor (GCSF) for poor prognosis acute myelogenous leukaemia patients. All patients in the trial receive the new treatment. The clinical endpoint is complete remission (CR) of the leukaemia. For patients achieving such a state, the treatment will be termed successful. The standard treatment is fludarabine + ara-C, for which the success rate is 50%. The use of GCSF would be considered beneficial if it increased the success rate to 70%.

2.1 Designs based on error rates

Phase III trials are usually designed based on consideration of error rates. A similar approach might be considered for the design of phase II trials. The outcome of a phase II trial is a decision whether or not the experimental therapy merits further investigation. The random nature of any data collected means that the wrong answer may be obtained (of course, it is not known that the answer is wrong at the time of the trial, and it may never be known). In imaginary repetitions, the trial will lead to such an error with a certain probability. The properties of the test can therefore be given in terms of its error rates: the probabilities of incorrectly concluding that the treatment is effective when it is not (the *type I error rate*) or incorrectly concluding that it is not effective when it actually is (the *type II error rate*). In terms of the example introduced above, the type I error rate is the probability of concluding that the use of GCSF is effective when the true success rate for the new treatment is actually 50%. The type II error rate is the probability of concluding that GCSF is not effective when the actual true success rate is 70%.

It is possible to calculate the sample size required for a trial to achieve specified error rates. Often the sample size required will vastly exceed the level of resources available for a phase II study. Specification of small error rates enables the sort of definitive analysis associated with a phase III trial, but also leads to correspondingly large sample sizes. If phase II studies are not to usurp the role of phase III trials, a different approach would appear to be more appropriate.

One strategy is to reduce the level of rigour required: that is, increase either the type I or type II error rate. Reduction of the sample size in phase III is sometimes achieved by allowing the type II error rate to be increased, that is reducing the power of the test. This reduces the

chance of detecting as significant a truly efficacious therapy. In phase II, such an approach may be less appropriate. Schoenfeld [3] points out that a better approach is to maintain the power and increase the type I error rate. This increases the risk of erroneously concluding that the treatment is worthy of further investigation but does not increase the risk of missing an efficacious treatment. In phase II, when the trial will be followed by further testing, a type II error may be more serious than a type I error. It is therefore Schoenfeld's suggestion that the type I error rate be increased from the usual 5%, possibly to as large a value as 25%.

Average sample sizes may also be reduced without increasing the error rates by the use of interim analyses. In studies of serious diseases such as cancer, there is an obvious need for designs which allow interim analyses of the data to be made so that the trial can be stopped as soon as possible if the new treatment is not superior to the current standard therapy. An approach for the correct use of interim analyses in phase III clinical trials was described in detail by Todd *et al.* [1] and similar methods can be applied in phase II trials. Sequential phase II trial designs based on the frequentist approach have been proposed and evaluated by a number of authors, including Fleming [4], Simon [5], Chen [6], and Conaway & Petroni [7]. Whilst the calculation of stopping rules so that the error rates are maintained involves complex statistical arguments and computation, the fact that the data may be summarized simply means that the resulting stopping rule is easily described.

An example of a design given by Simon [5] which might be appropriate for the GCSF example is given in Figure 1. For this design, the type I and type II error rates are both equal to 10% when the standard success rate is 50% and an improvement to 70% is being considered. The sample sizes and critical values for the two stages are carefully chosen to achieve these error rates, and Simon presents many similar designs for use in a variety of situations. The use of a design conducted in two stages means that the sample size is no longer fixed in advance, but depends on the data observed at the first stage, since early stopping is now possible. Although the maximum sample size is larger than that required by a fixed sample size trial with the same error rates, the chance to stop early means that on average the sample size is reduced.

A different approach, in which estimation as well as hypothesis testing is considered, was suggested by Gehan [8]. He also proposed that the trial be conducted in two stages. At the end of the first stage a decision is made to abandon development of the new treatment if there have been no treatment successes observed. The sample size for the first stage is determined so as to give a specified type I error rate. For a standard success rate of 50% and a required type I error rate of 5%, the first stage would include five

patients, although a more commonly used design is for a type I error rate of 5% and a standard success rate of 20%, for which the first stage includes 14 patients. Following the first stage, the second stage is planned depending on the data from the first stage, so as to estimate the unknown success rate for the new treatment with specified precision. More recently, this approach has been extended by Chen *et al.* [9] to trials conducted in three stages.

An alternative strategy is the use of a surrogate endpoint in phase II. As this endpoint may be less variable than the primary endpoint used in a phase III analysis, standard error rate constraints such as a 5% type I error rate and a 90% power can then be achieved with a relatively small sample size. The gain in power must, however, be balanced against the lack of information on the real question of interest when a decision is based on the surrogate rather than the primary endpoint. The type I error rate is the probability of proceeding to phase III when the effect of the experimental treatment on the primary endpoint (rather than on the surrogate) does not differ from that of the standard therapy. This will be above the nominal 5% level if there is less than perfect correlation between the primary and surrogate endpoints.

2.2 Bayesian designs

In the designs considered in Section 2.1, the sample size and the stopping rule are determined so as to give the correct answer with specified error rates. This approach seems appropriate in a phase III trial where a definitive conclusion is required. Phase II trials, however, are less formal. Information about the new therapy may be available from sources outside the trial. The incorporation of such information in phase III might be considered to bias the findings of the trial. In phase II, however, such information is likely to be used in decision-making processes whether or not it is incorporated formally in an analysis. The *Bayesian* approach allows initial belief about the true (unknown) success rate to be combined with the data observed. If interim analyses are performed, the belief is updated in the light of observed data as they accumulate. Initial belief is summarized by a *prior distribution* for the success rate that describes what values are most likely. For example, the upper graph in Figure 2 shows a possible prior distribution for the probability of success for the new treatment in the GCSF trial. Before the start of the trial we may have little evidence for believing the new treatment to be superior to the standard, so that values close to 0.5 (that is a 50% success rate) are most likely. The area of the shaded region to the right of the value 0.5 gives the probability that the true success rate is above 50%. Before the start of the experiment, this probability is equal to one half, indicating that we believe equally that the true rate could be above or below 50%. The flat shape of the

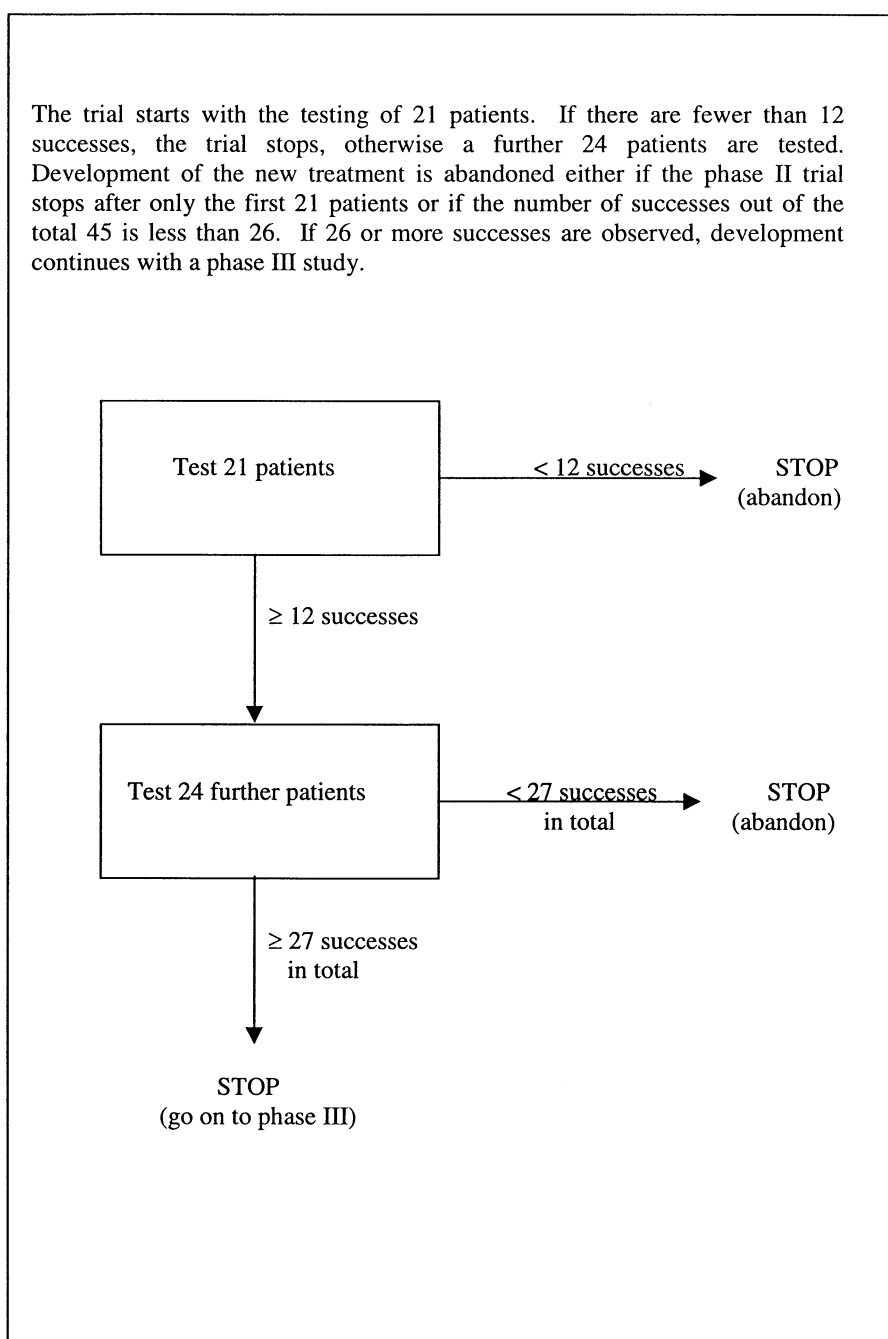


Figure 1 A two-stage design proposed by Simon [5].

curve also indicates our considerable uncertainty about the true success rate. After the observation of some data, the prior distribution is updated to give a *posterior distribution*. The lower graph in Figure 2 shows how the prior would be updated by observation of 7 successes out of 10 observations. The data are more positive than the initial belief described by the prior distribution, so that the posterior distribution indicates that values more than 0.5 are more likely. The probability that the success rate is above 50% is now increased, in this case to 0.86.

A method for the construction of a stopping rule for phase II trials based on the Bayesian approach has been proposed by Thall & Simon [2]. They suggest that the trial continues with regular interim analyses until belief about the true success rate is sufficiently precise to be able to say either that it is likely that the new treatment is better than the standard or that it is unlikely that the new treatment is better than the standard by a large enough margin to warrant further investigation. In the GCSF trial, the trial might be stopped as soon as the probability that the success

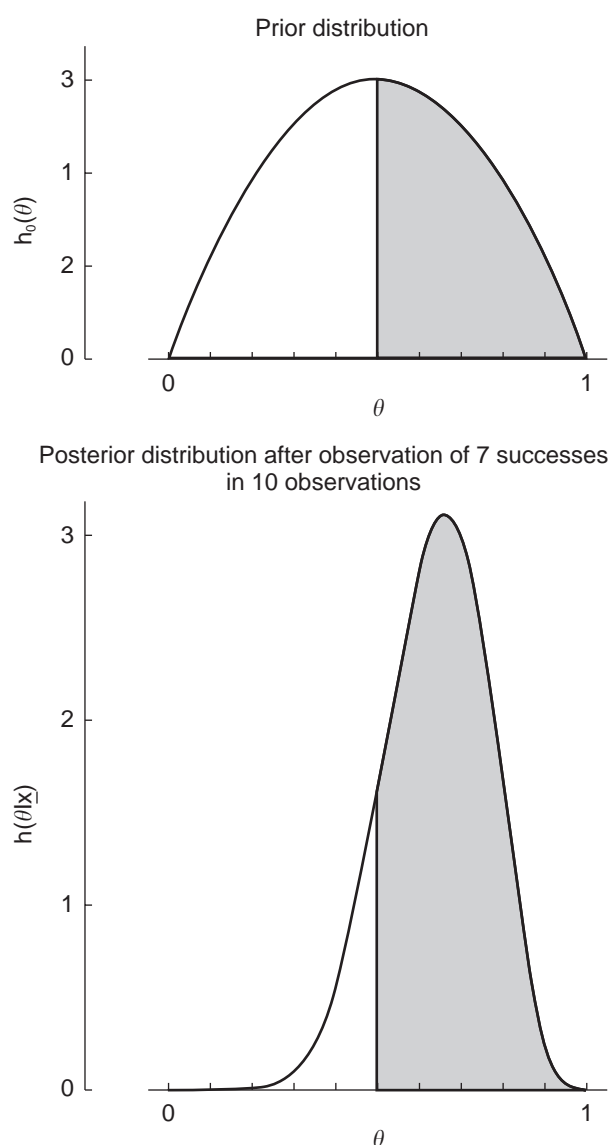


Figure 2 Prior and posterior distributions for the probability of success.

rate is at least 50% exceeds 0.9 (that is we are confident that the new treatment is at least as effective as the standard) or stopped for futility if the probability that the success rate is at least 70% is less than 0.1 (we are then confident that the new treatment does not represent a considerable improvement over the standard). In practice, by considering all possible data sets that might arise and determining when the trial would be stopped, the stopping rule may be described simply by rules of a similar form to that given in Figure 1.

2.3 Decision theory designs

Although it has had little application in practice, a third approach to the construction of stopping rules for single-arm studies is to base the decision on the possible

consequences. If, for example, the trial terminates and it is decided that development should continue with a phase III study, there is a risk that this decision is made erroneously, in which case there will be a cost associated with the time, resources and possibly suffering experienced by patients receiving an inferior treatment, associated with that decision. If a decision is made not to proceed to phase III, these costs will be avoided. Potential benefits associated with a truly superior treatment may be lost, however. The use of *Bayesian decision theory* enables belief about the true success rate to be coupled with an analysis of the costs and benefits associated with different actions to allow a rational decision to be made. At interim analyses, the decision as to whether or not the trial should stop may be made in a similar fashion by comparing costs associated with further testing with the gain in information given by additional data.

A number of authors have suggested designs based on such an approach (see, for example, Hilden *et al.* [10], Cressie & Biele [11], Heitjan [12], Stallard [13, 14], and Stallard *et al.* [15]). In spite of this work, such an approach remains little used. A practical barrier is the difficulty of specifying the consequences of all possible actions, and it is likely that this has prevented widespread use of this approach. It is worth noting that all of the stopping rules described in Sections 2.1 and 2.2 lead to a decision as to when to stop the trial and whether to go on with phase III testing. It is only the rules obtained using formal decision theory, however, that explicitly consider the consequences of that decision.

3 Comparative studies

Comparative studies, in which patients are randomized between a single experimental treatment and a control, which may either be a standard treatment used as an active control or a placebo, are almost universal in phase III. The use of a concurrent control group is important in removing bias due to systematic differences between historical control patients and those in the trial. Sample sizes in phase II are inevitably considerably smaller than those in phase III. This means that there is less scope for gaining information on both experimental and control patients and the division of patients between two groups reduces the precision of estimates obtained. In the presence of considerable historical knowledge of the standard treatment, data from a very small control group may provide relatively little additional information, so that a single-arm study may be a more efficient use of limited resources. It is not the purpose of the phase II trial to provide definitive evidence of treatment effect. The decision as to whether or not to include a control group depends on several factors such as the nature of the disease under investigation, the presence or absence of a standard

treatment, the number of patients available for the trial and the level of prior information about the properties of the control treatment.

In practice, phase II trials of cancer therapies are almost always single-arm studies and phase II trials in other indications are almost always comparative studies. In all areas a case-by-case consideration of why a control group is or is not being used might be more appropriate.

Although there has been little methodological work on stopping rules for comparative phase II studies, on the whole the principles underlying the approaches for single-arm studies can be extended to this setting. With two groups to consider, the stopping rules cannot be described as easily as in Figure 1, and in some cases the mathematical or computational detail can be considerably more burdensome than in the single-arm case. Following the structure of Section 2, and building on the concepts introduced there, we can consider stopping rules based on error rates, Bayesian methods and Bayesian decision theory.

The error-based designs for comparative phase III studies could obviously be used directly in the phase II setting. As in single arm trials, reduction of the sample size can be achieved by allowing the type I error rate to increase above its common 5% level. Allowing interim analyses again reduces the expected sample size without increasing error rates. Todd *et al.* [1] discuss the use of sequential analysis in phase III trials, and the methods considered there are also appropriate in the phase II setting if the error rates are increased to reduce the sample size below that usually considered in phase III. As described by Todd *et al.* this approach can also be used to monitor a continuous rather than a dichotomous endpoint.

The Bayesian approach discussed in Section 2.2 can also be easily extended to the case of a comparative study. In this case success rates on both experimental and control treatment are unknown. Rather than comparing the success rate on the experimental treatment with some fixed probability such as 0.5, the difference between success rates on the two arms can be considered. As an example, if a control treatment had also been included in the GCSF trial introduced above, a stopping rule might have been used in which the trial was stopped for superiority if the probability that the difference is larger than zero (that is that the experimental treatment is better than the control) is larger than 0.9, or for futility if the probability that the difference is less than 0.2 (that is that the experimental treatment is not a sufficient improvement over the control) is larger than 0.9.

The Bayesian approach can also be used to provide stopping rules for trials when a continuous endpoint is used. Some other measure of treatment difference will replace the difference in success rates and a distribution combining prior opinion and observed data will be

constructed for this measure. The decision as to when to stop the trial will then be based on this distribution.

There has been some work on the use of Bayesian decision theory in comparative studies, though this approach has had very little practical application. Designs for comparative studies with a binary endpoint have been developed by Lewis & Berry [16], whilst Berry & Ho [17] have developed designs for a continuous endpoint.

4 Selection studies

In a selection study, patients are assigned, usually at random, to one of a number of treatment groups. The different groups may receive different drugs or different doses or formulations of the same drug. A control treatment may or may not be included. Usually resources are only available to conduct a large phase III study on a very small number of new treatments, so that only the best one, or possibly two, treatments will be developed further.

As the trial progresses, less effective or safe treatments will be dropped from the study until a decision is finally made as to which, if any, of the remaining treatments warrant further investigation. Construction of a stopping rule for such a study is inevitably more complicated than for studies with a single experimental treatment.

Although it is possible to assess each treatment in turn, the resulting sample size is likely to be prohibitive. Thall *et al.* [18] suggested comparing the treatments in a single trial conducted in two stages. At the end of the first stage, a decision is made as to which treatment is the best. Only this treatment and the control then continue to the second stage. The dropping of all but the selected treatment and control at the interim analysis leads to a reduced sample size. An extension allowing further interim analyses in the second stage is described by Stallard & Todd [19].

The Bayesian method proposed by Thall & Simon [2] for single-arm studies has been extended to selection studies by Thall & Estey [20] and Thall & Sung [21] to provide a method in which ineffective treatment arms may be dropped early in the study.

Even using the approaches described here, the required sample size may be large. In the decision theoretic approach due to Whitehead [22, 23], the advantages of a thorough evaluation of a small number of experimental treatments are compared with the disadvantages associated with ignoring some treatments. Whitehead shows that to maximize the expected success rate for the treatment chosen to proceed to phase III it often makes sense to use sample sizes considerably smaller than those that would be used in more conventional designs.

5 Discussion

In contrast to phase III trials, phase II clinical trials are informal and exploratory in nature. Whether or not a

formal stopping rule is imposed at the design stage, the data will be monitored as they accumulate and investigation of an experimental therapy will cease if there is evidence of treatment-related adverse events or of a lack of efficacy. This paper has described some of the approaches available to formalize the decision as to when to stop a trial or drop a poorly performing therapy.

Specification of a stopping rule in advance is desirable and ensures that the conditions under which treatment should stop are considered before the trial starts. The exploratory nature of a phase II trial means, however, that it may be appropriate to alter or override the stopping rule during the course of the trial if unforeseen results are obtained. As a new treatment considered suitable for further testing will have to undergo the rigour of a phase III trial before accepting regulatory approval, the phase II trial can be slightly less formal.

Although many of the stopping rules described in the papers referenced in Sections 2–4 are based on complex statistical theory and extensive computation, they can often be summarized as simple rules, for example that shown in Figure 1. Here, the interim analyses comprise a simple counting of the number of successes and failures observed on each treatment arm. The depth of underlying theory is in contrast to the ease of application. The role of the clinical investigator is to decide on desirable properties for the trial, from which a statistician can find a suitable design. It is of little consequence to investigators how the stopping rule was devised, so long as it is seen that the recommendation to stop is made in appropriate circumstances.

The development of stopping rules in phase II has to date been mainly in the area of oncology, so that the methods proposed are most suitable for studies in serious diseases. The underlying concepts can be extended to application in other areas, leading to similar savings in efficiency. A collaboration between clinicians and statisticians is needed to aid the progress of such application.

The authors are grateful to two referees for their helpful comments on this paper.

References

- 1 Todd S, Whitehead A, Stallard N, Whitehead J. Interim analyses in phase III studies. *Br J Clin Pharmacol* 2001; **51**: 394–399.
- 2 Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**: 337–349.
- 3 Schoenfeld D. Statistical considerations for pilot studies. *Int J Radiation Oncol Biol Physics* 1980; **6**: 371–374.
- 4 Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**: 143–151.
- 5 Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials* 1989; **10**: 1–10.
- 6 Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Biometrics* 1997; **43**: 865–874.
- 7 Conaway M, Petroni G. Bivariate sequential designs for phase II trials. *Biometrics* 1995; **51**: 656–664.
- 8 Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis* 1961; **13**: 346–353.
- 9 Chen S, Soong S-J, Wheeler RH. An efficient multiple-stage procedure for phase II clinical trials that have high response rate objectives. *Controlled Clin Trials* 1994; **15**: 277–283.
- 10 Hilden J, Bock JE, Andreasson B, Visfeldt J. Ethics and decision theory in a clinical trial involving severe disfigurement. *Theoret Surg* 1987; **1**: 183–189.
- 11 Cressie N, Beale J. A sample-size-optimal Bayesian decision procedure for sequential pharmaceutical trials. *Biometrics* 1994; **50**: 700–711.
- 12 Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Statistics Med* 1997; **16**: 1791–1802.
- 13 Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998; **54**: 279–294.
- 14 Stallard N. Approximately optimal designs for phase II clinical studies. *J Biopharm Statistics* 1998; **8**: 469–487.
- 15 Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* 1999; **55**: 971–977.
- 16 Lewis RJ, Berry DA. Group-sequential clinical trials: a classical evaluation of Bayesian decision-theoretic designs. *J Am Statist Assoc* 1994; **89**: 1528–1534.
- 17 Berry DA, Ho C-H. One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics* 1988; **44**: 219–227.
- 18 Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* 1989; **45**: 537–547.
- 19 Stallard N, Todd SC. Sequential designs for phase III clinical trials incorporating treatment selection. Technical report 99/1, Department of Applied Statistics, The University of Reading.
- 20 Thall PF, Estey EH. A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation. *Statistics Med* 1993; **12**: 1197–1211.
- 21 Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics Med* 1998; **17**: 1563–1580.
- 22 Whitehead J. Designing phase II studies in the context of a programme of clinical research. *Biometrics* 1985; **41**: 373–383.
- 23 Whitehead J. Sample sizes for phase II and phase III clinical trials, an integrated approach. *Statistics Med* 1986; **5**: 459–464.